

Example 2: Identification of Novel Human Aspartyl Proteases Using Database Mining by Genome Bridging

Materials and Methods:

Computer-assisted analysis of EST databases, cDNA, and predicted polypeptide sequences:

Exhaustive homology searches of EST databases with the CEASP1, F21F8.3, F21F8.4, and F21F8.7 sequences failed to reveal any novel mammalian homologues. TBLASTN searches with R12H7.2 showed homology to cathepsin D, cathepsin E, pepsinogen A, pepsinogen C and renin, particularly around the DTG motif within the active site, but also failed to identify any additional novel mammalian aspartyl proteases. This indicates that the *C. elegans* genome probably contains only a single lysosomal aspartyl protease which in mammals is represented by a gene family that arose through duplication and consequent modification of an ancestral gene.

TBLASTN searches with T18H9.2, the remaining *C. elegans* sequence, identified several ESTs which assembled into a contig encoding a novel human aspartyl protease (Hu-ASP1). As is described above in Example 1, BLASTX search with the Hu-ASP1 contig against SWISS-PROT revealed that the active site motifs in the sequence aligned with the active sites of other aspartyl proteases. Exhaustive, repetitive rounds of BLASTN searches against LifeSeq, LifeSeqFL, and the public EST collections identified 102 EST from multiple cDNA libraries that assembled into a single contig. The 51 sequences in this contig found in public EST collections also have been assembled into a single contig (THC213329) by The Institute for Genome Research (TIGR). The TIGR annotation indicates that they failed to find any hits in the database for the contig. Note that the TIGR contig is the reverse complement of the LifeSeq contig that we assembled. BLASTN search of Hu-ASP1 against the rat and mouse EST sequences in ZooSeq revealed one homologous EST in each database (Incyte clone 700311523 and IMAGE clone 313341, GenBank accession number W10530, respectively).

TBLASTN searches with the assembled DNA sequence for Hu-ASP1 against both LifeSeqFL and the public EST databases identified a second, related human sequence (Hu-Asp2) represented by a single EST (2696295). Translation of this partial cDNA sequence reveals a single DTG motif which has homology to the active site motif of a bovine aspartyl protease, NM1.

BLAST searches, contig assemblies and multiple sequence alignments were performed using the bioinformatics tools provided with the LifeSeq, LifeSeqFL and LifeSeq Assembled databases from Incyte. Predicted protein motifs were identified using either the ProSite dictionary (Motifs in GCG 9) or the Pfam database.

#### Full-length cDNA cloning of Hu-Asp1

The open reading frame of *C. elegans* gene T18H9.2CE was used to query Incyte LifeSeq and LifeSeq-FL databases and a single electronic assembly referred to as 1863920CE1 was detected. The 5' most cDNA clone in this contig, 1863920, was obtained from Incyte and completely sequenced on both strands. Translation of the open reading frame contained within clone 1863920 revealed the presence of the duplicated aspartyl protease active site motif (DTG/DSG) but the 5' end was incomplete. The remainder of the Hu-Asp1 coding sequence was determined by 5' Marathon RACE analysis using a human placenta Marathon ready cDNA template (Clontech). A 3'-antisense oligonucleotide primer specific for the 5' end of clone 1863920 was paired with the 5'-sense primer specific for the Marathon ready cDNA synthetic adaptor in the PCR. Specific PCR products were directly sequenced by cycle sequencing and the resulting sequence assembled with the sequence of clone 1863920 to yield the complete coding sequence of Hu-Asp-1 (SEQ ID No. 1).

Several interesting features are present in the primary amino acid sequence of Hu-Asp1 (Figure 1, SEQ ID No. 2). The sequence contains a signal peptide (residues 1-20 in SEQ ID No. 2), a pro-segment, and a catalytic domain containing two copies of the aspartyl protease active site motif (DTG/DSG). The spacing between the first and second active site motifs is about 200 residues which should correspond to the expected size of a single, eukaryotic aspartyl protease domain. More interestingly, the sequence contains a predicted transmembrane domain (residues 469-492 in SEQ ID No.2) near its C-terminus which suggests that the protease is anchored in the membrane. This feature is not found in any other aspartyl protease.

#### Cloning of a full-length Hu-Asp-2 cDNAs:

As is described above in Example 1, genome wide scan of the *Caenorhabditis elegans* database WormPep12 for putative aspartyl proteases and subsequent mining of human EST databases revealed a human ortholog to the *C. elegans* gene T18H9.2 referred to as Hu-Asp1. The assembled contig for Hu-Asp1 was used to query for human paralogs using the BLAST search tool in human EST databases and a single significant match

(2696295CE1) with approximately 60% shared identity was found in the LifeSeq FL database. Similar queries of either gb105PubEST or the family of human databases available from TIGR did not identify similar EST clones. cDNA clone 2696295, identified by single pass sequence analysis from a human uterus cDNA library, was obtained from Incyte and completely sequenced on both strands. This clone contained an incomplete 1266 bp open-reading frame that encoded a 422 amino acid polypeptide but lacked an initiator ATG on the 5' end. Inspection of the predicted sequence revealed the presence of the duplicated aspartyl protease active site motif DTG/DSG, separated by 194 amino acid residues. Subsequent queries of later releases of the LifeSeq EST database identified an additional ESTs, sequenced from a human astrocyte cDNA library (4386993), that appeared to contain additional 5' sequence relative to clone 2696295. Clone 4386993 was obtained from Incyte and completely sequenced on both strands. Comparative analysis of clone 4386993 and clone 2696295 confirmed that clone 4386993 extended the open-reading frame by 31 amino acid residues including two in-frame translation initiation codons. Despite the presence of the two in-frame ATGs, no in-frame stop codon was observed upstream of the ATG indicating that the 4386993 may not be full-length. Furthermore, alignment of the sequences of clones 2696295 and 4386993 revealed a 75 base pair insertion in clone 2696295 relative to clone 4386993 that results in the insertion of 25 additional amino acid residues in 2696295. The remainder of the Hu-Asp2 coding sequence was determined by 5' Marathon RACE analysis using a human hippocampus Marathon ready cDNA template (Clontech). A 3'-antisense oligonucleotide primer specific for the shared 5'-region of clones 2696295 and 4386993 was paired with the 5'-sense primer specific for the Marathon ready cDNA synthetic adaptor in the PCR. Specific PCR products were directly sequenced by cycle sequencing and the resulting sequence assembled with the sequence of clones 2696295 and 4386993 to yield the complete coding sequence of Hu-Asp2(a) (SEQ ID No. 3) and Hu-Asp2(b) (SEQ ID No. 5), respectively.

Several interesting features are present in the primary amino acid sequence of Hu-Asp2(a) (Figure 2 and SEQ ID No. 4) and Hu-Asp-2(b) (Figure 3, SEQ ID No. 6). Both sequences contain a signal peptide (residues 1-21 in SEQ ID No. 4 and SEQ ID No. 6), a pro-segment, and a catalytic domain containing two copies of the aspartyl protease active site motif (DTG/DSG). The spacing between the first and second active site motifs is variable due to the 25 amino acid residue deletion in Hu-Asp-2(b) and consists of 168- versus-194 amino acid residues, for Hu-Asp2(b) and Hu-Asp-2(a), respectively. More